

ОТЗЫВ

научного руководителя
на выпускную квалификационную работу бакалавра
студентки кафедры системного программирования СПбГУ
Екатерины Юрьевны Леденевой
по теме «Сравнительный анализ алгоритмов вычисления текстовых
метрик для документации программного обеспечения»

Проблема разработки документации программного обеспечения (ПО) давно известна и широко обсуждается сообществом программистов и исследователей. В последнее время активно развиваются различные языки и инструменты для автоматизации разработки документации. Одним из них является JavaDoc, являющийся стандартом де-факто в сфере разработки Java-приложений. Данный подход позволяет создавать, в рамках определенных соглашений, комментарии к Java-коду прямо в исходных текстах программ и генерировать на их основе html и pdf документы. Однако для JavaDoc-комментариев остро стоят вопросы корректности, качества и соответствия коду. В настоящее время существует много исследовательских работ, посвященных анализу JavaDoc-комментариев.

Одним из направлений тут является повторное использование комментариев. В связи с этим имеется задача нахождения неточных повторов в комментариях. На настоящий момент не было исследований, рекомендующих алгоритмы и подходы, позволяющие это сделать. В имеющихся научных работах, как правило, используются средства, заимствованные из других областей, в частности, инструменты обнаружения клонов в программном коде.

В рамках данной работы проведено исследование известных строковых алгоритмов таких как Longest Common Sequence, Levinshtein Distance и др. на предмет применимости к вычислению сходства (pairwise similarity) JavaDoc-комментариев.

Екатерина Юрьевна полностью справилась с поставленной задачей. Ею было сделано следующее.

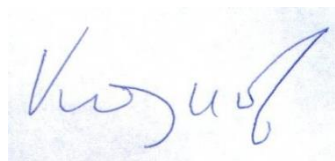
- Собран набор (dataset) из 2638 пар JavaDoc-комментариев размером, извлеченный из групп неточных повторов, созданных экспертами на основе известных open source Java-проектов GSON, JUnit4, Mockito,Slf4J.
- Спроектирован и реализован процесс обработки для вычисления сходства комментариев, состоящий из нормализации (лемматизация, удаление стоп-слов) и сегментации (потеговое сравнение комментариев).
- Подобраны пороги для классификации комментариев выбранными строковыми алгоритмами с помощью машинного обучения (логистической регрессии). Результаты оценены с помощью следующих метрик: F-мера, Accuracy, ROC AUC. Также была использована k-fold кроссвалидация.
- Выполнена оценка быстродействия алгоритмов.
- Инфраструктура эксперимента реализована на языке Python с использованием библиотек NLTK, SciPy, scikit-learn.

Екатерина Юрьевна получила интересные научные результаты, которые будут в скором времени опубликованы.

Следует отметить самостоятельность и исполнительность Екатерины Юрьевны. А также свободное владение как современными концепциями и алгоритмами Data Science, так и превосходное владение навыками практического программирования. Фактически, Екатерина Юрьевна является сформировавшимся специалистом в области Data Science, а также талантливым исследователем. Я считаю её одной из самых лучших студенток за всю историю кафедры системного программирования.

Рекомендую оценку «отлично».

Доктор технических наук,
профессор кафедры системного
программирования СПбГУ, доцент



/Д.В.Кознов/